GENDER-INDEPENDENT SPEAKER RECOGNITION USING SOURCE NORMALISATION

Mitchell McLaren and David A. van Leeuwen

Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands

{m.mclaren, d.vanleeuwen}@let.ru.nl

ABSTRACT

Source-normalisation (SN) was proposed to improve the robustness of i-vector-based speaker recognition for under-resourced and unseen cross-speech-source evaluation conditions. The technique of source-normalisation estimates directions of undesired within-speaker variation more accurately than traditional methods when cross-source variation is not explicitly observed from each speaker in system development data. Incorporated into Within Class Covariance Normalisation (WCCN), source-normalisation provides significant improvements to speaker recognition based on i-vectors.

This paper proposes a novel approach to gender-independent Probabilistic LDA (PLDA) through the use of SN-WCCN to normalise for the variation that separates genders as a pre-processing step for i-vector based PLDA classification. Evaluated on the NIST 2010 speaker recognition evaluation (SRE) dataset, the proposed approach demonstrated performance comparable to a typical gender-dependent configuration.

***Index Terms*—** gender-independent speaker recognition, probabilistic linear discriminant analysis, i-vectors

1. INTRODUCTION

The goal of gender-independence in text-independent speaker recognition systems is to make less assumptions about the sex of speakers in a trial while pertaining to the same performance as gender-dependent (GD) systems. The difficultly in attaining this goal has resulted in GD configurations becoming commonplace in recent NIST speaker recognition evaluations (SRE) [1]. Gender-dependent systems assume that the sex of the speaker in an audio recording is known. In a forensics scenario, however, this gender label cannot always be accurately determined. Desired is a gender-independent (GI) speaker recognition system without gender-labelling or gender-detection at recognition time while offering similar performance to the GD alternative.

The authors of [2] proposed a GI approach based on state-of-the-art PLDA speaker recognition using i-vectors as features. This system combined scores from gender-conditioned PLDA classifiers based on the gender likelihood of each i-vector in a trial. While this configuration does not require that a speaker's sex be labelled at recognition time, the combination of scores relies on the accuracy of a gender-classifier. A GI system void of gender classification would be preferable in forensic scenarios and cases where limited computational power is available. Recently, the source-normalised approach to LDA [3, 4] was quite successful at normalizing the i-vector distribution between speech sources, and we were therefore motivated to see if speaker sex can be treated as a kind of 'source variation' that can be reduced via source-normalisation.

This research was funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 238803.

Source-Normalised LDA and Within-Class Covariance Normalisation (SN-LDA and SN-WCCN, respectively) were originally proposed to suppress directions from the i-vector space that were due to the differences in speech source (ie., telephone and microphone speech) [3, 4]. One of the key aspects of source-normalisation is the manner in which within-speaker variation is estimated from a training dataset that does not include utterances from each source of interest from every speaker — an implicit requirement of traditional LDA and WCCN algorithms [4]. Intuitively, it can be seen that *gender* is another source of variation in which a speaker can not provide examples for each source of interest and is potentially detrimental to gender-independent speaker recognition.

In this work we extend source-normalisation to the task of gender-independent speaker recognition by labelling gender as the source to be normalised. We employ SN-WCCN as a preprocessing step to i-vector-based PLDA classification. The presence of gender and speech source variation in the i-vector space is illustrated graphically along with ability of SN-WCCN to suppress this variation. Based on this analysis, it is hypothesised that SN-WCCN can be utilised to suppress the directions contributing to gender separation to realise an improvement in gender-independent speaker recognition. The proposed technique is compared to gender-independent recognition and the approach of Senoussaoui *et al.* [2] using the recent NIST 2010 SRE.

This paper is structured as follows. Section 2 describes the PLDA i-vector framework for speaker recognition. Section 3 provides details on existing techniques for gender-independence in PLDA classification and the SN-WCCN algorithm used in this work. The experimental protocol and corresponding results are given in Sections 5 and Section 6.

2. SPEAKER RECOGNITION USING I-VECTORS

Over the past two years, the i-vector approach to speaker recognition [5] has become the standard in this research field, showing top performance in comparative evaluation studies [1]. This work focuses on the PLDA model for i-vector classification in which state-of-the-art technology continues to advance rapidly.

2.1. I-vector extraction

I-vectors are a compact representation of an utterance having a dimensionality between that of feature space and the GMM supervector space. I-vectors are extracted from a total variability subspace $\boldsymbol{T}$ trained via factor analysis [6] such that it bounds the main directions of between-utterance variability. The total variability subspace is trained under the assumption that an utterance can be represented by the Gaussian Mixture Model (GMM) mean supervector,

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}, \tag{1}$$

where $\boldsymbol{M}$ consists of a speaker- and session-independent mean supervector $\boldsymbol{m}$ from the Universal Background Model (UBM) and a mean offset $\boldsymbol{Tw}$. The low-rank vector $\boldsymbol{w}$ (400 dimensions in this work) has a standard normal distribution $\mathcal{N}(0,1)$ and is referred to as the *i-vector*. An i-vector is essentially a maximum-a-posteriori point estimate in the space defined by $\boldsymbol{T}$ based on the observed statistics from an utterance. Readers are directed to [6] and [5] for detailed information on subspace training and extraction of i-vectors.

2.2. Transforming the i-vector space

Recent work has highlighted a number of steps to enhance the i-vector space which allows traditional PLDA to achieve similar performance to the more complex heavy-tailed PLDA [7]. These processes include WCCN [8], followed by the normalisation of i-vector length [9].

Within-Class Covariance Normalisation (WCCN) aims to normalise the within-speaker variance in i-vector space. The WCCN transform, $\boldsymbol{B}$, is derived through the Cholesky decomposition of $\boldsymbol{W}^{-1} = \boldsymbol{B}\boldsymbol{B}^t$ where the within-speaker covariance matrix

$$\boldsymbol{W} = \frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{N_s}(\boldsymbol{w}_i^s - \boldsymbol{\mu}_s)(\boldsymbol{w}_i^s - \boldsymbol{\mu}_s)^t. \quad (2)$$

Here, S is the number of speakers that each provide N_s i-vectors in the training dataset, and the i-vector mean from speaker s is equated as $\boldsymbol{\mu}_s = \frac{1}{N_s}\sum_{i=1}^{N_s}\boldsymbol{w}_i^s$.

I-vector length normalisation is the simple technique of normalising i-vectors to unit length. This process effectively Gaussianises the i-vector distribution so as to better fit to the Gaussian assumptions made by the PLDA model. The mapping of i-vector $\boldsymbol{w}$ to WCCN-projected and length-normalised $\overline{\boldsymbol{w}}$ then becomes, $\overline{\boldsymbol{w}} = \frac{\boldsymbol{B}^t\boldsymbol{w}}{|\boldsymbol{B}^t\boldsymbol{w}|}$.

2.3. Probabilistic Linear Discriminant Analysis (PLDA)

Probabilistic Linear Discriminant Analysis (PLDA) is a probabilistic approach to determine the likelihood that two i-vectors involved in a trial originate from the same speaker (i.e., a target trial). PLDA assumes that i-vector $\boldsymbol{w}_s$ from speaker s can be modelled by,

$$\boldsymbol{w}_s = \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{U}\boldsymbol{x} + \boldsymbol{\epsilon} \quad (3)$$

where subspaces $\boldsymbol{V}$ and $\boldsymbol{U}$ represent the major directions in which speaker and session variation reside, respectively. Factors $\boldsymbol{y}_s$ and $\boldsymbol{x}$ are the loading factors for the speaker and session subspaces and have a standard normal distribution while $\boldsymbol{\epsilon}$ represents the residual variation with diagonal covariance $\boldsymbol{\Sigma}$ matrix of dimension D. In this work, $D = 400$ and both $\boldsymbol{V}$ and $\boldsymbol{U}$ are subspaces of 200 dimensions.

A likelihood ratio for the comparison of two i-vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ is given by the ratio of two hypotheses: H_{tar} in which both i-vectors originate from the same speaker and the alternate hypothesis H_{non}. The likelihood ratio, R, is given by,

$$R = \frac{P(\boldsymbol{w}_1, \boldsymbol{w}_2|H_{\text{tar}})}{P(\boldsymbol{w}_1|H_{\text{non}})P(\boldsymbol{w}_2|H_{\text{non}})} \quad (4)$$

where the likelihoods $P(\boldsymbol{w}_1, \boldsymbol{w}_2|H_{\text{tar}})$ and $P(\boldsymbol{w}|H_{\text{non}})$ follow from the PLDA model using the approach outlined in [7].

3. DEALING WITH GENDERS IN PLDA

This section details the three different PLDA configurations analysed in Section 6. The recently proposed approach of Senoussaoui *et al.* [2] is presented along with traditional GD and GI systems.

3.1. Gender-dependent PLDA (GD-PLDA)

A gender-dependent (GD) i-vector framework typically extracts i-vectors from a gender-conditioned total variability subspace and employs gender-specific processes for both system development and classification. This configuration requires gender-labelled utterances at recognition time and is commonly employed in the NIST SREs [1]. Results from GD-PLDA can be viewed as the performance objective of a gender-independent system.

3.2. Gender-independent PLDA (GI-PLDA)

The gender-independent configuration in this work differs from GD-PLDA in that i-vectors are extracted from a *gender-independent* total variability subspace. We refer to GI-PLDA as a system that pools development data from both genders in all system processes and does not require gender-labelled data.

3.3. Gender-mixture PLDA (Gmix-PLDA)

The gender-independent approach to speaker recognition in [2] is based on a mixture of PLDA models and operates in the following manner. A development set of i-vectors, extracted from the same subspace as used for GI-PLDA, are used to train male- and female-conditioned PLDA classifiers. Comparison of i-vectors involves firstly obtaining likelihood ratios R_M and R_F from the gender-conditioned PLDA systems. A gender discrimination score, G_i, is then calculated for each i-vector using a gender-classifier $G_i = \frac{P(\boldsymbol{w}_i|M)}{P(\boldsymbol{w}_i|F)}$. These scores can be obtained directly from the PLDA model as in [2]. Following equations (8–10) of [2] and substituting equiprobable trial-type and gender priors, scores are then combined to produce the final gender-independent likelihood ratio,

$$R = \frac{2(R_M G_1 G_2 + R_F)}{(G_1 + 1)(G_2 + 1)}. \quad (5)$$

For the remainder of this document this configuration will be referred to as gender-mixture PLDA (Gmix-PLDA).

Equation (5) can be viewed as a weighting of gender-conditioned scores which is dependent on gender classification score G_i. As with most classifiers, the estimate of G_i will incur some error. In contrast to the 2% EER of G_i in [2], our system (as detailed in Section 5) obtained an EER of 5% with which we were unable to surpass performance of GI-PLDA. For comparative purposes, we present results for a Gmix-PLDA (Oracle) configuration which assigns G_i based on known gender labels. This is indicative of the *gender-dependent* performance obtained by the authors in [2] which was found to be attainable through their implementation of Gmix-PLDA. The following section proposes to extend source-normalised WCCN to provide an alternative approach to achieving gender-independence without the need for gender classification.

4. GENDER-NORMALISATION

This section proposes the use of Source-Normalised WCCN to suppress the variation that separates genders in the i-vector space to improve the traditional GI-PLDA configuration.

4.1. Source normalisation

Source normalisation was proposed in the context of LDA in [3] to improve the robustness of i-vector based speaker recognition to cross speech source comparisons when using a suboptimal LDA training dataset. A suboptimal dataset was defined as one absent of i-vectors

from each source of interest (such as microphone and telephone speech sources) from every speaker. This common type of dataset was shown to adversely effect the within-speaker covariance matrix calculated in the traditional manner using (2). Source normalisation is a technique to better estimate between- and within-speaker scatter matrices in this context.

Source normalisation involves firstly estimating the between-speaker scatter, $\boldsymbol{S}_B^{\text{src}}$, for each source of interest, src, by assuming the source mean $\boldsymbol{\mu}_{\text{src}}$ is the center of the i-vector space. That is,

$$\boldsymbol{S}_B^{\text{src}} = \sum_{s=1}^{S_{\text{src}}} N_s(\boldsymbol{\mu}_s - \boldsymbol{\mu}_{\text{src}})(\boldsymbol{\mu}_s - \boldsymbol{\mu}_{\text{src}})^t, \tag{6}$$

where $\boldsymbol{\mu}_{\text{src}} = \frac{1}{N_{\text{src}}} \sum_{n=1}^{N_{\text{src}}} \overline{\boldsymbol{w}}_n^{\text{src}}$ and N_{src} designates the number of speech samples taken from source src. The final between-speaker scatter is the accumulation of these between-speaker scatter matrices $\overline{\boldsymbol{S}}_B = \sum \boldsymbol{S}_B^{\text{src}}$. The within-speaker covariance matrix is defined as the residual variation from the total covariance matrix, $\boldsymbol{S}_T = \sum_{n=1}^{N} \overline{\boldsymbol{w}}_n \overline{\boldsymbol{w}}_n^t$, such that,

$$\boldsymbol{S}_W = \boldsymbol{S}_T - \overline{\boldsymbol{S}}_B. \tag{7}$$

4.2. Gender-Normalised WCCN (GN-WCCN)

In this work, we exploit source normalisation in the WCCN pre-processing phase of the GI-PLDA configuration to realise an improved gender-independent classifier. Source-normalised WCCN (SN-WCCN) was originally shown to be comparable to SN-LDA in [4] when normalising for differences between speech sources in the traditional i-vector framework. SN-WCCN involves utilising $\boldsymbol{W} = \frac{1}{S}\boldsymbol{S}_W$ instead of (2) in the Cholesky decomposition to obtain WCCN transform $\boldsymbol{B}$.

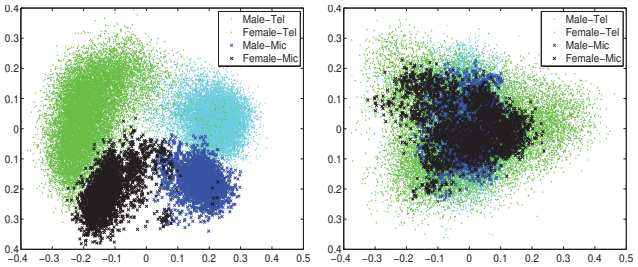
In the context of gender-independent classification, the source to be normalised is the gender — that is, $\text{src} = \{\text{male}, \text{female}\}$. We term this process *Gender-Normalised* WCCN (GN-WCCN). GN-WCCN is incorporated into the GI-PLDA system such that gender-labels are required only to estimate of the within-speaker scatter during system development and is gender-blind at trial time thus eliminating the need for gender-classification.

We also investigate the additional normalisation of speech-source along with gender. Microphone and telephone speech sources are considered. Referred to as Gender- and Speech-source-Normalised WCCN (GSN-WCCN), four sources are assigned in this scenario: $\text{src} = \{\text{male}_{\text{tel}}, \text{male}_{\text{mic}}, \text{female}_{\text{tel}}, \text{female}_{\text{mic}}\}$.

5. EXPERIMENTAL PROTOCOL

The recent NIST 2010 SRE corpus was used to evaluate the proposed techniques. Results are reported for four evaluation conditions corresponding to det conditions 2–5 in the evaluation plan. These include *int-int*, *int-mic*, *int-tel*, and *tel-tel* trials from the *extended* protocols [1]. Performance was evaluated using the equal error rate (EER) and a normalised minimum decision cost function (DCF) calculated using $C_{\text{M}} = 1$, $C_{\text{FA}} = 1$ and $P_{\text{T}} = 0.001$.

Speech activity detection was performed as in [4]. GI and GD 2048-component UBMs were trained on 20-dimensional, feature-warped MFCCs (including C_0) with deltas and double-deltas appended. Development data was sourced from the NIST 2004—2006 SRE corpora and LDC releases of Fisher English, Switchboard II: phase 3 and Switchboard Cellular (parts 1 and 2). The same data was utilised during total variability subspace training while Fisher data was excluded for WCCN and PLDA training. Utterances from each speaker were sourced from either microphone or telephone speech.



(a) WCCN (b) GSN-WCCN

Fig. 1. Projection of i-vectors into 2D PCA space after applying WCCN and GSN-WCCN. This represents the gender- and source-separation in i-vector space prior to training the PLDA classifier.

6. RESULTS

The following experiments illustrate the effectiveness of the proposed gender-normalised approach to PLDA by comparing it to the alternate configurations detailed in Section 3.2. The effect of normalising for speech-source in each configuration is also analysed.

6.1. Analysis of gender-normalised i-vector space

The effect of source normalisation in reducing the separation between i-vectors from different speech sources was first illustrated in [4]. We illustrate in this work that the separation between genders can be removed from i-vector space prior to training the PLDA classifier. Figure 1 depicts the distribution of male and female i-vectors from both microphone and telephone speech sources after applying WCCN or GSN-WCCN, length normalisation, and projection into a 2D-PCA space. As expected, the greatest observable variation in Figure 1(a) when using WCCN is the separation between speaker sex. The differences between microphone- and telephone-sourced i-vectors are also clearly evident. Figure 1(b) illustrates that the application of GSN-WCCN effectively removed variation from the i-vector space that was attributed to both speech-source and gender. Although restricted to two dimensions in the figure, GSN-WCCN appears to provide a more homogeneous distribution of i-vectors. The effect of this process is analysed in terms of recognition performance in the following section.

6.2. Gender normalisation

The top half of Table 1 presents the SRE'10 results using gender-independent (GI) and gender-mixture (Gmix) PLDA and the proposed use of gender-normalised (GN) WCCN prior to GI-PLDA. Each of these systems utilise i-vectors extracted from a gender-independent total variability subspace and do not require gender-labelling at recognition time. Results from the gender-dependent (GD) system are also detailed in the table where it can be seen that this system offered the best performance in the majority of cases. It should be noted that the GD configuration in Table 1 differs from the GD system in [2] which utilised i-vectors extracted from a GI subspace and is equivalent to Gmix (Oracle) in this work.

Gmix-PLDA (Oracle) was found to offer some improvement over GI-PLDA in cross-channel conditions (*int-tel* and *int-mic*), however no benefit was observed in the alternate conditions with the exception of improved EER in *tel-tel* trials. Interestingly, the application of GN-WCCN instead of WCCN prior to GI-PLDA resulted in only marginal relative improvements and yet a notable performance improvement over Gmix (Oracle) in the *int-int* trials. This suggests that suppressing the variation attributed to gender alone via

		int-int		int-tel		int-mic		tel-tel	
WCCN Type	**PLDA Type**	Min. DCF	EER	Min. DCF	EER	Min. DCF	EER	Min. DCF	EER
WCCN	GI	.6031	3.66%	.6004	3.56%	.4644	2.72%	**.5027**	3.11%
WCCN	Gmix (Oracle)	.6833	3.76%	.5876	3.21%	.4724	2.56%	.5275	**2.70%**
GN-WCCN	GI	**.5905**	3.46%	.5913	3.18%	.4598	2.64%	.5030	2.91%
WCCN	GD	.6466	**3.45%**	**.5462**	**2.98%**	**.4420**	**2.40%**	.5043	2.78%
SN-WCCN	GI	.5543	3.43%	.5166	2.85%	.4404	2.58%	.5037	3.11%
SN-WCCN	Gmix (Oracle)	.6472	3.45%	.5194	2.81%	.4513	2.42%	.5229	2.71%
GSN-WCCN	GI	**.4931**	**2.93%**	.4928	**2.71%**	**.3739**	2.28%	.5067	2.96%
SN-WCCN	GD	.5715	3.00%	**.4650**	2.71%	.4081	**2.14%**	**.4958**	**2.76%**

Table 1. SRE'10 (extended) results comparing gender-independent (GI), gender-mix (Gmix) and gender-dependent (GD) PLDA to the proposed GN-WCCN preprocessing phase for GI-PLDA with and without speech-source-normalisation (SN).

GN-WCCN provided no substantial benefit to classification performance of the gender-independent configuration. It was observed in Section 6.1 that speech-source variation was the next major source of variation in the i-vector space. The following section investigates whether normalising both gender and speech-source variation via GSN-WCCN can better realise a gender-normalised configuration.

6.3. Normalisation of gender- and speech-source

Speech-source normalisation was incorporated into each system configuration by utilising either SN-WCCN instead of WCCN for pre-processing or, in the case of the proposed GI-PLDA configuration, Gender- and Speech-source-Normalised (GSN-WCCN) as detailed in Section 4.2. It was hypothesised that removing speech-source variation alongside gender variation would close the gap between GD-PLDA and the proposed GI configuration. SRE'10 results are reported in the bottom half of Table 1. Comparing these results to those in the top half of Table 1, it can be observed that performance metrics from all PLDA configurations were improved through the normalisation of speech source with the exception of *tel-tel* trials which were largely unaffected. This reflects previous findings when using SN-LDA [4]. The Gmix (Oracle) configuration offered comparable performance to GI-PLDA when the i-vector space was pre-processed with SN-WCCN. In contrast, GSN-WCCN offered considerable benefit to the GI-PLDA system such that performance was largely comparable to GD-PLDA after SN-WCCN. In some cases, the proposed system outperformed GD-PLDA with the most notable relative improvement of 14% in minimum DCF of the *int-int* trials. It is hypothesised that the under-resourced interview conditions benefited most from the two-fold increase in system development data allowing source-normalisation to better estimate the directions of speech source variation common to both genders.

7. DISCUSSION

The experiments in Section 6 demonstrated that GI-PLDA can be improved by incorporating gender- and speech-source-normalisation in the WCCN processing phase prior to PLDA modelling. On one hand, this improvement could be expected due to a better estimate of the within-speaker covariance to be normalised via WCCN from a training dataset absent of cross-gender variation. On the other hand, speaker sex may be viewed as a valuable source of between-speaker variation as it describes a characteristic of the speaker. Interestingly, normalising for gender alone via GN-WCCN provided marginal improvements over GI-PLDA. When combined with speech-source normalisation, however, gender normalisation was found to offer considerable benefit. One hypothesis for this system behaviour is that speech source variation is more detrimental to classification performance than gender variation. In removing gender variation from the i-vector space, directions of speech-source variation common to both genders could be more accurately estimated due to the increase in development data for under-resourced speech in a gender-pooled dataset, thus resulting in improved system performance.

8. CONCLUSION

A novel gender-independent (GI) PLDA classifier for i-vector based speaker recognition was proposed and evaluated. Gender-normalised WCCN (GN-WCCN) was employed as a pre-processing step prior to GI-PLDA training and classification. Evaluated on the recent NIST 2010 SRE, GN-WCCN provided marginal benefits in terms of performance. The concept of gender-normalisation was better realised, however, by additionally normalising for speech-source variation via GSN-WCCN with which performance metrics were largely comparable to the commonly utilised gender-dependent (GD) configuration. Under-resourced trial conditions found particular benefit from GSN-WCCN due to the two-fold increase in system development data from pooled gender datasets.

9. REFERENCES

[1] National Institute of Standards and Technology, *NIST Speaker Recognition Evaluation site*, Available: http://www.itl.nist.gov/iad/mig/tests/sre/.

[2] M. Senoussaoui, P. Kenny, N. Brümmer, E. de Villiers, and P. Dumouchel, "Mixture of PLDA models in i-vector space for gender-independent speaker recognition," in *Proc. Interspeech*, 2011, pp. 25–28.

[3] M. McLaren and D.A. van Leeuwen, "Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors," in *Proc. IEEE ICASSP*, 2011, pp. 5456–5459.

[4] M. McLaren and D.A. van Leeuwen, "Source-normalised LDA for robust speaker recognition using i-vectors," *In print, IEEE Trans. Audio Speech and Language Processing*, 2011.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, pp. 788–798, 2011.

[6] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, , and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, pp. 980–988, 2008.

[7] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brümmer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *Proc. IEEE ICASSP*, 2011.

[8] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. Ninth Int. Conf. on Spoken Language Processing*, 2006, pp. 1471–1474.

[9] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, 2011, pp. 249–252.